บทที่ 2

ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

ในปัจจุบันนี้การสื่อสารแบบใช้สายโดยใช้เส้นใยแก้วนำแสง(optical fiber)ที่มีขนาดของแถบความถี่สูงเริ่มมีราคาลดลงและมีการใช้งานกันมากขึ้น แต่อย่างไรก็ตามในการสื่อสารแบบไร้สายเช่นโทรศัพท์เคลื่อนที่แบบเซลลูลาร์(cellular) หรือการสื่อสารดาวเทียมก็ยังคงมีความจำเป็นในการสงวนการใช้แถบความถี่อยู่ สัญญาณเสียงพูดที่ใช้กันในปัจจุบันส่วนใหญ่จะอยู่ในรูปแบบของสัญญาณดิจิตอล เพื่อที่จะได้ทำการประมวลผล เก็บรักษาหรือทำการส่งได้โดยใช้โปรแกรมคอมพิวเตอร์ การเก็บข้อมูลแบบดิจิตอลทำให้เกิดอัตราของข้อมูลที่สูง ซึ่งต้องการแถบความถี่ในการส่งหรือเก็บที่สูงเช่นกัน "การเข้ารหัสเสียงพูด(Speech coding)หรือการบีบอัดเสียงพูด(Speech Compression)"คือวิธีการหารูปแบบของข้อมูลดิจิตอลที่กะทัดรัดเพื่อใช้แสดงแทนสัญญาณเสียงพูดเพื่อวัตถุประสงค์ในการส่งและเก็บอย่างมีประสิทธิภาพ หลักการสำคัญคือการนำเสนอเสียงพูดโดยใช้จำนวนบิตที่น้อยที่สุดในขณะที่ยังคงรักษาคุณภาพในการรับฟังของเสียงนั้นเอาไว้ได้

การเข้ารหัสเสียงเกี่ยวข้องกับการสุ่มข้อมูล (sampling) และการควอนไตซ์ขนาด (Amplitude quantization) แต่ขนาดอัตราสุ่มข้อมูลของเสียงพูดมักจะคงที่ประมาณ 2 เท่าของขนาดแถบความถี่ของสัญญาณเสียงแบบอนาลอก ดังนั้นการควอนไตซ์จะเป็นวิธีการสำคัญในการกำหนดรูปแบบการเข้ารหัสเสียงพูด ซึ่งแบ่งได้เป็น 2 แบบใหญ่ ๆ คือ

1. การควอนไตซ์โดยตรงหรือแบบนอนพาราเมตริก คือการนำสัญญาณไบนารีมาแสดงแทนสัญญาณเสียงโดยตรง

2. การควอนไตซ์แบบพาราเมตริก คือการนำสัญญาณไบนารีมาแสดงแทนโมเดลของเสียงหรือ/และพารามิเตอร์ทางความถี่ของเสียง

วิธีการควอนไตซ์ที่สำคัญจะกล่าวถึงในหัวข้อถัดไป

เสียงพูดโดยปกติของมนุษย์มีความถี่ไม่เกิน 4 กิโลเฮิรตซ์และจะถูกสุ่มข้อมูลที่อัตรา 8 กิโลเฮิรตซ์ การเข้ารหัสเสียงโดยการควอนไตซ์โดยตรงที่ง่ายที่สุดคือ Pulse -Code Modulation (PCM) เสียงพูดที่เข้ารหัสที่อัตรา 64 กิโลบิตต่อวินาทีที่ใช้ logarithm PCM ถูกเรียกว่าการเข้ารหัสเสียงแบบ "ไม่บีบอัดข้อมูล(noncompressed)" ใช้เป็นการเข้ารหัสแบบอ้างอิงเทียบกับการเข้ารหัสแบบอื่น ๆ ซึ่งอาจจะแบ่งการเข้ารหัสเสียงพูดเป็นพวกได้ตามอัตราข้อมูลได้ดังนี้

1. การเข้ารหัสอัตราข้อมูลสูง ใช้อัตราข้อมูลมากกว่า 16 กิโลบิตต่อวินาที
2. การเข้ารหัสอัตราข้อมูลขนาดกลาง ใช้อัตราข้อมูลอยู่ในช่วง 8-16 กิโลบิตต่อวินาที
3. การเข้ารหัสอัตราข้อมูลต่ำ ใช้อัตราข้อมูลอยู่ในช่วง 2.4-8 กิโลบิตต่อวินาที
4. การเข้ารหัสอัตราข้อมูลต่ำมาก ใช้อัตราข้อมูลต่ำกว่า 2.4 กิโลบิตต่อวินาที

การเข้ารหัสเสียงพูดที่อัตราตั้งแต่ขนาดกลางลงมานั้นต้องใช้ขบวนการวิเคราะห์-สังเคราะห์(analysis-synthesis) ในขั้นของการวิเคราะห์จะหาชุดของพารามิเตอร์ที่ใช้แทนสัญญาณเสียงที่ถูกเข้ารหัสได้อย่างมีประสิทธิภาพ และในขั้นของการสังเคราะห์ค่าพารามิเตอร์เหล่านี้จะถูกถอดรหัสและสร้างเสียงพูดกลับมา การวิเคราะห์อาจจะเป็นได้ทั้งแบบวงปิด(closed loop)หรือแบบวงเปิด(open loop) ในแบบวงปิดค่าพารามิเตอร์จะถูกค้นหาจากค่าความแตกต่างระหว่างเสียงต้นฉบับกับเสียงที่ถูกสร้างขึ้นมา นั่นคือในส่วนวงปิดจะต้องมีส่วนสังเคราะห์อยู่ภายใน กระบวนการแบบหลังนี้จะเรียกว่าการวิเคราะห์จากการสังเคราะห์(analysis - by - synthesis)

การเข้ารหัสเสียงพูดแบบพาราเมตริกอาจเรียกได้อีกอย่างว่า "การเข้ารหัสเสียงตามลักษณะของเสียงพูดหรือโวโคดเดอร์(Speech-specific coders หรือ Voice coders หรือ Vocoders)" เป็นการเข้ารหัสที่เน้นในเรื่องของคุณภาพใน

การรับฟังของเสียงพูดโดยไม่จำเป็นต้องได้สัญญาณที่เหมือนเดิมทุกประการ โวโคดเดอร์สามารถทำงานที่อัตราข้อมูลต่ำมากได้โดยให้คุณภาพของเสียงที่ระดับเสียงสังเคราะห์ โดยที่ในอัตราข้อมูลที่สูงขึ้นก็จะให้คุณภาพของเสียงที่ดีขึ้น

2.1 การวัดสมรรถนะ (performance)

สมรรถนะในการเข้ารหัสเสียงจะพิจารณาคุณภาพจากคุณสมบัติต่าง ๆ เช่น อัตราข้อมูล(bit rate) คุณภาพของเสียงที่ถูกสร้างกลับมา ความซับซ้อนของอัลกอริธึม ช่วงเวลาประวิง(delayed time)และความทนทานต่อความผิดพลาดภายในช่องสัญญาณ(Channel errors) หรือการแทรกสอดที่เกิดจากเสียงสะท้อน(Acoustic interferences) โดยปกติแล้วการเข้ารหัสเสียงที่อัตราข้อมูลต่ำแต่ให้คุณภาพของเสียงที่สูง สามารถทำได้โดยใช้อัลกอริธึมที่มีความซับซ้อนสูง สำหรับตัวอย่างการทำงานตามเวลาจริงของการเข้ารหัสแบบ CELP ซึ่งจะกล่าวถึงในหัวข้อถัดไปนั้นมีคุณสมบัติดังกล่าวจะต้องการการคำนวณเป็นจำนวนหลายล้านคำสั่งต่อ 1 วินาที(Million-Instruction Per Second หรือ MIPS)

ในการสื่อสารข้อมูลแบบดิจิตอล คุณภาพของเสียงพูดถูกแบ่งออกเป็น 4 ระดับชั้นทั่ว ๆ ไปคือ ระดับกระจายเสียง(broadcast),ระดับเครือข่าย(network หรือ toll) ,ระดับการสื่อสาร(communications)และสุดท้ายคือระดับสังเคราะห์ (synthetic)

1.เสียงพูดระดับกระจายเสียง เสียงพูดนี้ใช้อ้างอิงถึงเสียงพูดบรรยายคุณภาพสูงซึ่งสามารถสร้างขึ้นโดยใช้อัตราข้อมูลที่สูงกว่า 64 กิโลบิตต่อวินาทีขึ้นไป

2. เสียงพูดคุณภาพระดับเครือข่าย ให้คุณภาพที่สามารถเปรียบเทียบได้กับเสียงพูดแบบอะนาลอกความถี่ระหว่าง 200-3200 เฮิรตซ์และสามารถสร้างได้โดยใช้อัตราข้อมูลมากกว่า 16 กิโลบิตต่อวินาที

3. เสียงพูดระดับสื่อสาร มีการลดทอนของคุณภาพลงไป ทำให้ดูไม่เป็นธรรมชาติ แต่สามารถเข้าใจได้โดยง่าย และมีคุณภาพเพียงพอที่จะใช้ในการสื่อสาร สามารถสร้างได้โดยอัตราข้อมูลมากกว่า 4.8 กิโลบิตต่อวินาทีและมีเป้าหมายที่ 4 กิโลบิตต่อวินาที

4. เสียงพูดระดับสังเคราะห์ ปกติแล้วสามารถรับฟังได้เข้าใจแต่ไม่เป็นธรรมชาติและสูญเสียคุณสมบัติในการรู้จำ(recognition)เจ้าของเสียงพูด ได้จากอัตราข้อมูลน้อยกว่า 4 กิโลบิตต่อวินาที

การวัดคุณภาพของเสียงเป็นงานที่สำคัญและมีความยากอยู่มาก วิธีการหนึ่งที่นิยมใช้กันคืออัตราส่วนกำลังสัญญาณต่อกำลังของเสียงรบกวน(Signal-to-Noise Ratio หรือ SNR) เป็นวิธีการแบบเชิงวัตถุ(objective measurement)ใช้สำหรับวัดคุณสมบัติของอัลกอริธึมที่ใช้ในการบีบอัดข้อมูล โดยมีการคำนวณดังสมการ (2.1)

$$SNR = 10\log_{10}\left\{\frac{\sum_{n=0}^{M-1} s^2(n)}{\sum_{n=0}^{M-1}\left(s(n)-\hat{s}(n)\right)^2}\right\} \qquad (2.1)$$

โดยที่ $s(n)$ คือ เสียงต้นฉบับ

$\hat{s}(n)$ คือ เสียงที่ผ่านการเข้ารหัสและถอดรหัสออกมา

SNR เป็นการวัดความถูกต้องของการสร้างเสียงกลับมาแบบช่วงยาว (long-term) ซึ่งมีแนวโน้มที่จะซ่อนเสียงรบกวนที่เกิดขึ้นเพียงชั่วคราวระหว่างการสร้างเสียงกลับมา โดยเฉพาะในสัญญาณเสียงที่มีขนาดเล็ก

การเปลี่ยนแปลงอย่างฉับพลันสามารถตรวจจับ และประเมินได้โดยการใช้ SNR ในช่วงสั้นนั่นคือการคำนวณ SNR สำหรับแต่ละส่วนของเสียงพูดที่มีอยู่ N จุด การวัดที่จะแสดงให้เห็นถึงจุดที่อ่อนแอของสัญญาณเสียงคือ การทำ SNR ทีละส่วน(segmental SNR หรือ SEGSNR) โดยมีการคำนวณดังนี้

$$SEGSNR = \frac{10}{L}\sum_{i=0}^{L-1}\log_{10}\left\{\frac{\sum_{n=0}^{N-1}s^2(iN+n)}{\sum_{n=0}^{N-1}\left(s(iN+n)-\hat{s}(iN+n)\right)^2}\right\} \qquad (2.2)$$

เนื่องจากการหาค่าเฉลี่ยของสมการ (2.2) เกิดหลังการคำนวณค่าลอการิธึม SEGSNR จะแสดงข้อผิดพลาดของตัวเข้ารหัสที่การทำงานมีการเปลี่ยนแปลงไปเรื่อย ๆ ได้มากกว่า SNR ธรรมดา นอกจากนี้ยังมีวิธีการอื่น ๆ อีกเช่น articulation index , the log special distance และ euclide distance ซึ่งวิธีการเหล่านี้เป็นการวัดทางวัตถุทั้งสิ้น ไม่ได้พิจารณาถึงคุณภาพในการรับฟังเสียงของหูมนุษย์ แต่ในการออกแบบอัลกอริธึมที่ใช้อัตราข้อมูลต่ำเกือบทั้งหมดจะมีพื้นฐานมาจากบรรทัดฐานของการรับฟังเสียงของมนุษย์

วิธีการในการวัดคุณภาพเสียงอีกวิธีหนึ่งคือ 'การวัดในเชิงของผู้ฟัง(subjective measurement)' เช่น Diagnostic Rhyme Test (DRT) , Diagnostic Acceptability Measure (DAM) และ Mean Opinion Score (MOS) ซึ่งมีพื้นฐานบนการให้คะแนนของผู้ฟังซึ่งได้รับการฝึกหัดมาโดยเฉพาะ

MOS เป็นวิธีที่ใช้กันมากที่สุดในการกำหนดคุณภาพของเสียงพูด ซึ่งจะใช้ผู้ฟังประมาณ 12-14 คน ผู้ซึ่งได้รับการฝึกหัดให้สามารถให้คะแนนเสียงที่ได้รับการอัดแบบ Phonetics สมดุลเป็นระดับ 5 ระดับ แสดงในตารางที่ 2.1 ระดับ 5 แสดงถึงเสียงที่ไม่แตกต่างจากเสียงต้นฉบับและไม่มีเสียงรบกวนที่สามารถรู้สึกได้ ส่วนระดับ 1 แสดงถึงเสียงที่มีเสียงรบกวนจำนวนมากและเสียงที่ไม่เหมือนจริง

ในการทดสอบแบบ MOS ผู้รับฟังจะต้องถูกปรับเทียบ (calibrated) ในกรณีที่ผู้รับฟังมีความคุ้นเคยกับสภาพในการรับฟังและระดับของเสียงที่พวกเขาได้รับฟัง การกำหนดคะแนนจะถูกเฉลี่ยจากการทดสอบการรับฟังตัวอย่างเสียงที่ผ่านการเข้ารหัสแล้วจำนวนหลายร้อยเสียง

Table 1 The MOS Scale

MOS Scale	Speech quality
1	bad
2	poor
3	fair
4	good
5	excellent

ตารางที่ 2.1 ระดับคะแนนของการทดสอบแบบ MOS

เสียงคุณภาพต่าง ๆ ได้คะแนน MOS ดังนี้ คะแนน 4-4.5 แสดงถึงคุณภาพของเสียงระดับเครือข่าย คะแนน 3.5-4.0 แสดงถึงคุณภาพระดับการสื่อสาร และคะแนน 2.5-3.5 แสดงถึงเสียงระดับเสียงสังเคราะห์ แต่การทดสอบ

แบบ MOS อาจจะเปลี่ยนแปรไปได้เสมอจากการทดสอบต่างครั้งกัน ดังนั้นจึงไม่สามารถเปรียบเทียบคุณภาพของการเข้ารหัสที่ใช้วิธีการที่แตกต่างกันได้อย่างสมบูรณ์

การทดสอบแบบขึ้นกับผู้ฟังแบบนี้ใช้เวลานานและสิ้นเปลืองค่าใช้จ่ายมาก ในปัจจุบันกำลังมีการพัฒนาการวัดคุณภาพของการเข้ารหัสเสียงแบบขึ้นกับผู้ฟังที่สามารถทำงานได้ด้วยตัวเอง หรือการวัดแบบเชิงวัตถุที่สามารถใช้ทำนายคุณภาพจากการวัดแบบขึ้นกับผู้ฟังได้

2.2 การควอนไตซ์แบบเวกเตอร์(Vector Quantization หรือ VQ) [7]

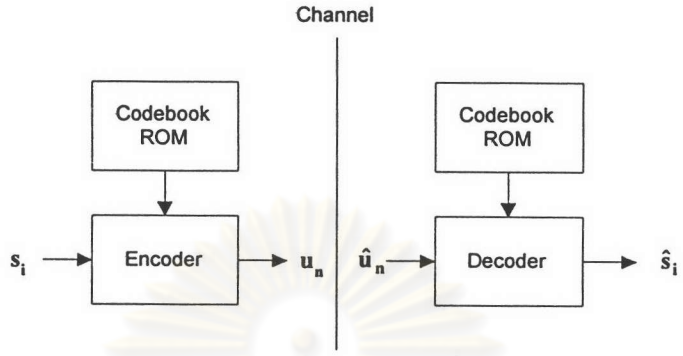
ชนิดของการเข้ารหัสสัญญาณจะเป็นชนิดพาราเมตริกหรือนอนพาราเมตริกขึ้นอยู่กับว่าการเข้ารหัสนั้นควอนไตซ์ตัวสัญญาณจริงหรือค่าพารามิเตอร์ของสัญญาณ การควอนไตซ์แบ่งได้เป็น 2 ประเภทใหญ่ ๆ คือการควอนไตซ์แบบสเกลาร์และการควอนไตซ์แบบเวกเตอร์ การควอนไตซ์แบบเวกเตอร์มีการศึกษาเป็นอย่างมากในช่วง 15 ปีที่ผ่านมา โดยแสดงถึงศักยภาพสูงในการเข้ารหัสเสียงพูดที่อัตราข้อมูลต่ำและให้คุณภาพเสียงที่สูง

1) การควอนไตซ์แบบสเกลาร์ (Scalar Quantization) คือการแทนค่าสัญญาณอนาลอก 1 ตัวด้วยชุดของสัญญาณไบนารี 1 ชุด สัญญาณไบนารีชุดหนึ่งจะแทนข้อมูลได้ N ระดับ ค่าของ N ขึ้นอยู่กับจำนวนบิตที่ใช้ L โดย $N = 2^L$ ถ้าใช้จำนวนบิตมากขึ้นจำนวนระดับก็จะมากขึ้นและเพิ่มความถูกต้องของการควอนไตซ์ แต่อัตราข้อมูลที่ใช้ก็จะเพิ่มมากขึ้นด้วย ผลต่างของระดับของการควอนไตซ์ที่อยู่ติดกันเรียกว่าขนาดขั้น(Step size)ขนาดขั้นที่ใช้อาจจะคงที่หรือไม่คงที่ขึ้นอยู่กับรายละเอียดของการควอนไตซ์ ตัวอย่างเช่น นอนอะแดปทีฟยูนิฟอร์ม PCM จะมีขนาดขั้นที่คงที่ส่งผลให้อัตราข้อมูลสูง ขนาดขั้นที่ไม่คงที่มีใช้ในนอนยูนิฟอร์ม PCM นั่นคือใช้ขนาดขั้นที่ละเอียด(Fine step size)ในช่วงของขนาดที่สัญญาณเกิดบ่อยและใช้ขนาดขั้นที่หยาบ(Coarse step size)ในช่วงของขนาดที่สัญญาณเกิดน้อย การออกแบบขนาดขั้นอาจจะใช้รูปร่างของฟังก์ชันความหนาแน่นของความน่าจะเป็น(Probability Density Function หรือ PDF) ของสัญญาณก็ได้ การออกแบบขนาดขั้นแบบลอการิทึมมีใช้ในนอนยูนิฟอร์ม PCM ที่เรียกว่า μ-law และ A-law ใช้การควอนไตซ์แบบลอการิทึมด้วยสัญญาณไบนารีขนาด 7 บิตทำให้ได้คุณภาพของเสียงเทียบเท่ากับการควอนไตซ์แบบยูนิฟอร์มด้วยสัญญาณไบนารีขนาด 12 บิต นอกจากนี้ยังมีขนาดขั้นที่สามารถปรับเปลี่ยนไปตามรูปแบบของสัญญาณที่เข้ามาได้ด้วย ขนาดขั้นแบบนี้มีใช้ในอะแดปทีฟ PCM (APCM) และยังมีการเข้ารหัสอื่นๆอีกหลายวิธีที่ใช้การจัดการกับขนาดขั้นแบบต่าง ๆตามที่ได้กล่าวมาแล้วได้แก่ ดิฟเฟอเรนเชียล PCM (DPCM) เดลตามอดูเลชั่น (DM) อะแดปทีฟ DPCM (ADPCM) อะแดปทีฟ DM (ADM)

2) การควอนไตซ์แบบเวกเตอร์(VQ) การบีบอัดข้อมูลด้วย VQ ทำได้โดยการเข้ารหัสชุดของข้อมูลอนาลอกในรูปแบบของบล็อกหรือเวกเตอร์ นั่นคือการแทนที่สัญญาณอนาลอกหลายตัวด้วยชุดของสัญญาณไบนารี 1 ชุด ถึงแม้ว่าความเข้าใจในข้อดีของ VQ ที่มีเหนือการควอนไตซ์แบบสเกลาร์จะมีมานานแล้ว แต่การเข้ารหัสเสียงที่ใช้วิธีของ VQ จริง ๆ ยังไม่มีจนกระทั่งทศวรรษที่ 1970 ทั้งนี้เป็นเพราะความซับซ้อนในการคำนวณของ VQ นั่นเอง ในปัจจุบันมีวิธีการที่มีประสิทธิภาพในการเข้ารหัสข้อมูลที่เป็นบล็อกหลายมิติทำให้สามารถใช้งาน VQ เพื่อการเข้ารหัสเสียงพูดคุณภาพสูงที่อัตราข้อมูลต่ำได้

การทำงานของ VQ ประกอบด้วยตัวควอนไตซ์แบบ N มิติและบล็อกของชุดรหัส (codebook) เวกเตอร์ที่เข้ามาจะถูกสร้างขึ้นจากตัวอย่างสุ่มของสัญญาณที่อยู่ติดกันหรือจากค่าพารามิเตอร์ของสัญญาณเสียงนั้น ตัวควอนไตซ์จะทำการจับคู่เวกเตอร์ขนาด Nx1 ที่เข้ามาตัวที่ i $s_i=[s_i(1)\ s_i(2)\ \ldots s_i(N)]$ เข้ากับสัญลักษณ์ที่จะถูกส่งไปในช่องสัญญาณ (channel symbol) $\{u_n\ ,n=1,2,\ldots,L\}$ สำหรับกรณีทั่วไปเราถือว่าในช่องสัญญาณไม่มีเสียงรบกวนดังนั้น $u_n = \hat{u}_n$ ในชุดรหัสจะ

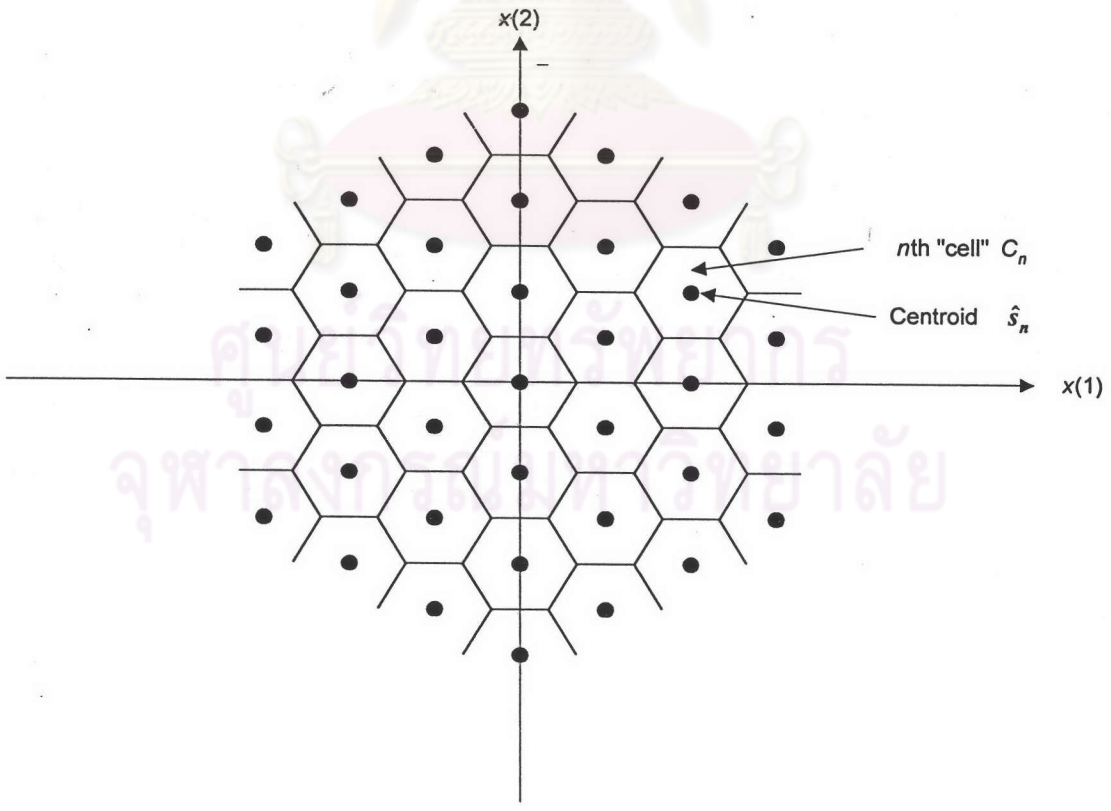
ประกอบด้วยเวกเตอร์รหัส (codevector) L ตัว $\{\hat{s}_n = [\hat{s}_n(1)\hat{s}_n(2)\ldots\hat{s}_n(N)]^T, n = 1,2,\ldots,L\}$ ซึ่งอยู่ในหน่วยความจำของทั้งตัวรับและตัวส่ง แสดงดังในรูปที่ (2.1)



รูป 2.1. แผนภาพการทำงานของการควอนไตซ์แบบเวกเตอร์

วิธีการค้นหาเวกเตอร์ที่ต้องการของ VQ ทำงานดังนี้ เวกเตอร์ s_i ที่เข้ามาจะถูกเปรียบเทียบกับเวกเตอร์รหัสทีละตัวทุก ๆ ตัว และค่าดัชนีที่ใช้แทนเวกเตอร์รหัสที่ใกล้เคียงกับ s_i ที่สุดจะถูกส่งไปในช่องสัญญาณ การพิจารณาความใกล้เคียงของเวกเตอร์รหัสจะดูจากการวัดความเพี้ยน(distortion measure) $\in (s_i, \hat{s}_n)$ วิธีการที่ใช้มากที่สุดและพื้นฐานที่สุดในการวัดความเพี้ยนคือการหาค่าผลรวมของความผิดพลาดยกกำลังสอง(sum of square error)ดังในสมการ

$$\in (s_i, \hat{s}_n) = \sum_{k=1}^{N} (s_i(k) - \hat{s}_i(n))^2 \qquad (2.3)$$



รูป 2.2 ตัวอย่างของการควอนไตซ์ในปริภูมิ 2 มิติ

เวกเตอร์รหัสจำนวน L ตัวที่อยู่ในชุดรหัสคือเวกเตอร์ค่าจริงขนาด Nx1 จำนวน L ตัวถูกออกแบบโดยการแบ่งปริภูมิเวกเตอร์ออกเป็นเซลที่ไม่ซ้อนทับกัน(nonoverlapping cell) จำนวน L เซล C_n แสดงดังในรูป 2.2 แต่ละเซล C_n จะถูกเชื่อมโยงกับเวกเตอร์ $\hat{s}_n$ ตัวควอนไตซ์จะระบุสัญลักษณ์ของช่องสัญญาณ u_n ให้ s_i เมื่อ s_i อยู่ใน C_n นั่นคือเมื่อ s_i อยู่ใน C_n มันจะถูกนำเสนอโดยให้ค่าเป็น $\hat{s}_n$ ซึ่งเป็นจุดศูนย์กลาง(centroid)ของเซล C_n สัญลักษณ์ของช่องสัญญาณมักจะใช้เป็นสัญญาณไบนารีของดัชนีหรือตำแหน่งของ $\hat{s}_n$ รูปแบบที่ง่ายที่สุดของ VQ คือเวกเตอร์ PCM (VPCM) ซึ่งใช้การค้นหาชุดรหัสแบบเต็มรูปแบบ(fully search) เรียกอีกอย่างหนึ่งว่า full search VQ หรือ F-VQ

จำนวนบิตต่อสัญญาณคือ $B = (\log_2 L)/N$ (2.4)

VQ มีข้อเสียเปรียบคือเรื่องความซับซ้อนสูงในการค้นหาชุดรหัส วิธีการบางวิธีในการลดความซับซ้อนจะทำให้ลดคุณภาพของเสียงในการเข้ารหัสลงหรือเพิ่มขนาดของหน่วยความจำที่ต้องใช้ วิธีลดอีกทางหนึ่งคือการทำให้ชุดรหัสเป็นบรรทัดฐานเดียวกัน(normalized) และมีการเข้ารหัสอัตราขยายแยกออกไปส่วนหนึ่ง เทคนิกนี้เรียกว่า VQ แบบอัตราขยาย/รูปร่าง(Gain/Shape VQ หรือ GS-VQ)นำเสนอโดย Buso *et. al.* รูปร่างของสัญญาณจะใช้เวกเตอร์รหัสจากชุดรหัสที่เป็นรูปร่าง ส่วนอัตราขยายหาได้จากชุดรหัสอัตราขยาย GS-VQ จะให้คุณภาพเสียงที่ดีกว่า F-VQ ที่มีความซับซ้อนเท่ากันอยู่ประมาณ 0.7 เดซิเบล ซึ่ง GS-VQ นี้ใช้กันอย่างกว้างขวางในการเข้ารหัสแบบ CELP รวมทั้ง LD-CELP ที่จะกล่าวถึงในหัวข้อถัดไป นอกจากนี้ยังมี VQ แบบที่ปรับเปลี่ยนได้หรืออะแดปทีฟ VQ (A-VQ) ซึ่งใช้ในการเข้ารหัส VSELP ซึ่งรายละเอียดของการทำงานจะไม่กล่าวถึงในที่นี้

2.3 ทฤษฎีการเข้ารหัสแบบการทำนายพันธะเชิงเส้น

การเข้ารหัสแบบการทำนายพันธะเชิงเส้น(Linear Predictive Coding หรือ LPC)ใช้หลักการของแบบจำลองการทำนายพันธะเชิงเส้น(Linear Predictive model หรือ LP model)เป็นระบบที่คล้ายคลึงกับรูปแบบการกำเนิดเสียงของมนุษย์ซึ่งมีส่วนประกอบที่สำคัญ 2 ส่วนคือต้นกำเนิดเสียงและทางเดินเสียง

ต้นกำเนิดเสียง มี 2 ลักษณะ คือ

1. เสียงก้อง(voiced sound) มีลักษณะเป็นรายคาบ(periodic)และมีค่าคาบของเสียง(pitch period)เป็นส่วนประกอบสำคัญ ตัวอย่างของเสียงก้องได้แก่ เสียงสระต่างๆ เสียงพยัญชนะ บ,ก ซึ่งเปล่งออกมาทางปาก เสียงพยัญชนะ น,ม,ง ซึ่งเปล่งออกมาทางจมูกที่เรียกว่าเสียงนาสิก(Nasal sound)

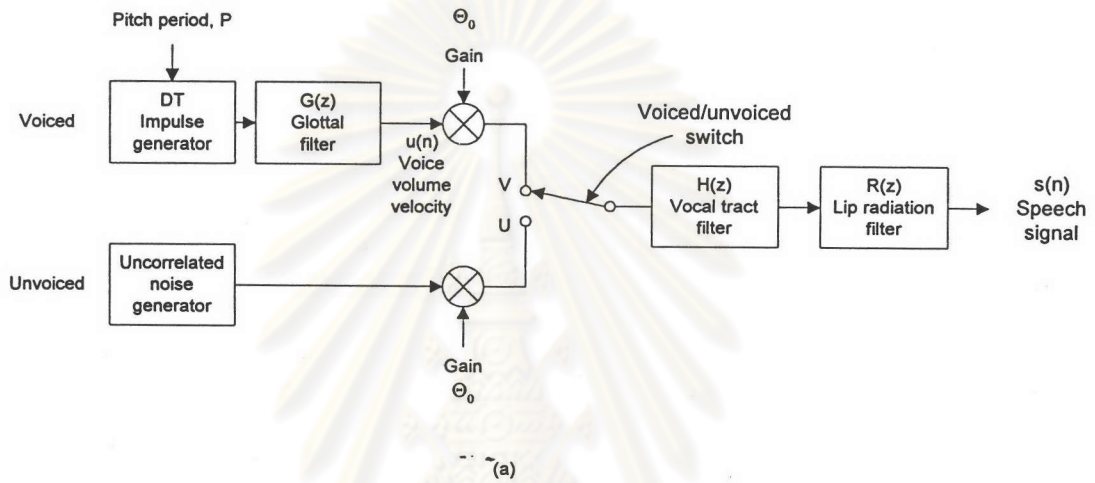
2. เสียงไม่ก้อง(unvoiced sound) ไม่เป็นรายคาบ แต่จะมีลักษณะเป็นสัญญาณของเสียงรบกวน(noise) ตัวอย่างของเสียงไม่ก้องได้แก่ เสียงพยัญชนะ ฟ,ส,ช

ทางเดินเสียง(Vocal tract)

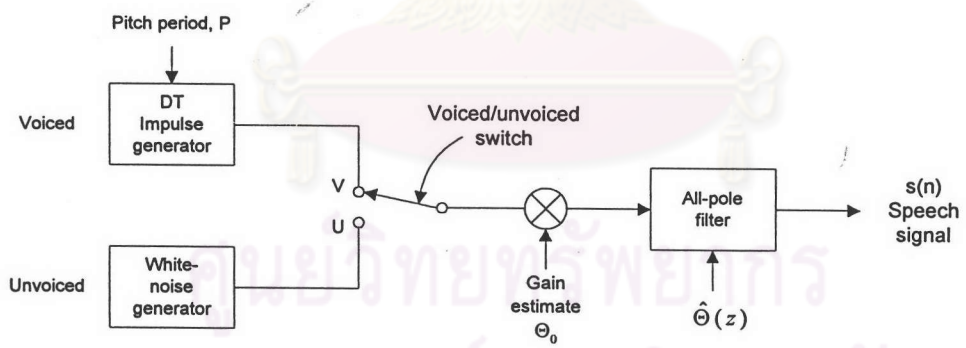
คือช่องที่เสียงจะเดินทางผ่านจากช่องกำเนิดเสียงถึงริมฝีปาก แสดงถึงส่วนประกอบของอวัยวะผลิตเสียง ส่วนของทางเดินเสียงคือช่องระหว่างลิ้นไก่(valum)เพดานแข็ง(hard palate)กับลิ้น(tonque) โดยเสียงทั้งเสียงก้องและไม่ก้องก็ต้องผ่านช่องทางนี้ทั้งสิ้น ส่วนในกรณีของเสียงนาสิกเสียงจากต้นกำเนิดจะต้องผ่านทางโพรงจมูก(nasal cavity)

ควบคู่ไปกับผ่านทางปาก จากรูปแบบการกำเนิดเสียงจริงของมนุษย์สามารถเขียนเป็นแบบจำลองได้ดังในรูป 2.3(a) และ LPC จะมีแบบจำลองดังในรูป 2.3(b)

$$\Theta(z) = \Theta'_0 \frac{1+\sum_{i=1}^{L} b(i)z^{-i}}{1+\sum_{i=1}^{R} a(i)z^{-i}} \qquad (2.5)$$



(a)



(b)

รูป 2.3. (a) แบบจำลองการกำเนิดเสียงจริง (b) แบบจำลองเลียนแบบการกำเนิดเสียงจริงโดยใช้การวิเคราะห์แบบการทำนายพันธะเชิงเส้น(LPC)

ฟังก์ชันถ่ายโอนของแบบจำลองการกำเนิดเสียงจริงในช่วงขณะที่เสียงมีคุณสมบัติ stationary อาจแทนได้ด้วยสมการ (2.5) เป็นวงจรกรองที่มีทั้งโพลและศูนย์ (Pole-zero filter) ส่วนฟังก์ชันถ่ายโอนของ LPC แสดงในสมการ (2.6)

โดยที่ $$\hat{\Theta}(z) = \frac{1}{1-\sum_{i=1}^{M}\hat{a}(i)z^{-i}} \quad (2.6)$$

สาเหตุสำคัญอย่างหนึ่งที่ LPC ใช้ตัวกรองที่มีแต่โพล(All-pole filter)แทนที่จะใช้ตัวกรองที่มีทั้งโพลและศูนย์ (Pole-zero filter)ก็เพราะตัวกรองที่มีแต่โพลจะใช้เฉพาะค่าของเอาท์พุทตัวเก่าๆในการคำนวณโดยใช้ค่าของอินพุทน้อยที่สุดโดยใช้เฉพาะตัวปัจจุบัน

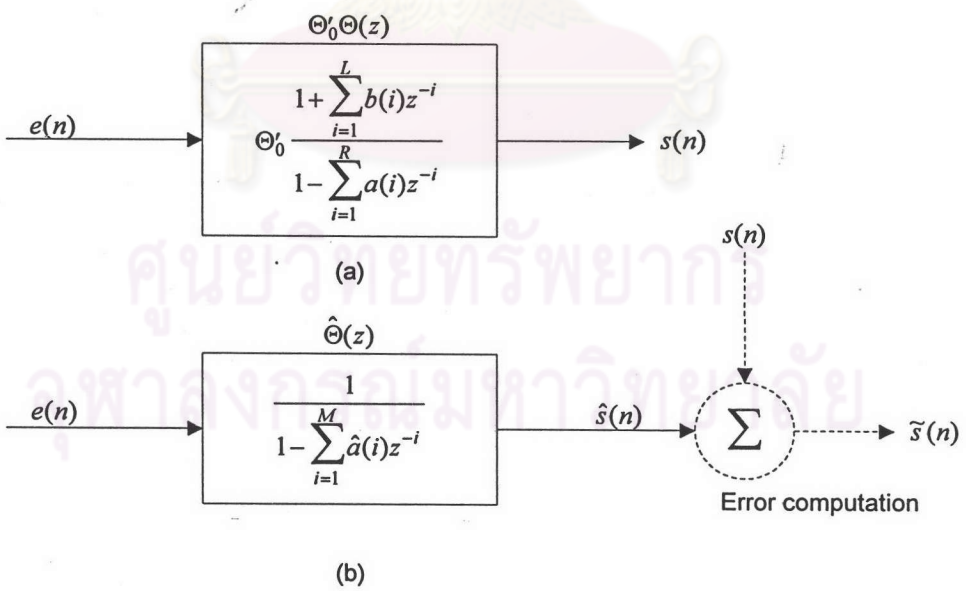
ถ้าให้ $Y(z)$ เป็นผลลัพท์และ $X(z)$ เป็นข้อมูลขาเข้าของฟิลเตอร์ จะได้

$$Y(z) = X(z) + \hat{a}(1)z^{-1}Y(z) + \hat{a}(2)z^{-2}Y(z) + \ldots + \hat{a}(M)z^{-M}Y(z) \quad (2.7)$$

และ $$y(n) = x(n) + \hat{a}(1)y(n-1) + \hat{a}(2)y(n-2) + \ldots + \hat{a}(M)y(n-M) \quad (2.8)$$

นั่นคือสามารถทำนายค่า $y(n)$ จากผลบวกเชิงเส้นของ $x(n)$ และ $y(n-i)$; $i = 1,2,\ldots,M$ เป็นที่มาของชื่อการทำนายพันธะเชิงเส้น ค่าพารามิเตอร์ $\hat{a}(i)$ จะถูกเรียกว่าสัมประสิทธ์ของการทำนาย(Prediction coefficients)

การเข้ารหัสแบบ LPC จะใช้ฟิลเตอร์ซึ่งมีลักษณะเป็นส่วนกลับของ $\hat{\Theta}(z)$ นั่นคือ $\hat{A}(z) = \hat{\Theta}^{-1}(z)$ มีอินพุทคือข้อมูลเสียง $s(n)$ และได้เอาท์พุทออกมาเป็น $\hat{e}(n)$ เรียกว่าลำดับของตัวกระตุ้น(Exitation sequence)เปรียบได้กับเสียงจากต้นกำเนิดเสียงของมนุษย์ มีทั้งชนิดที่เป็นรายคาบและชนิดที่ไม่เป็นรายคาบ(เสียงรบกวน)ซึ่งลำดับของตัวกระตุ้นนี้จะมีขนาดของข้อมูลที่ลดลงไปมากเมื่อเทียบกับข้อมูลเสียงต้นฉบับ ส่วนการถอดรหัสจะใช้ $\hat{e}(n)$ เป็นอินพุทของฟิลเตอร์ $\hat{\Theta}(z)$ เพื่อสร้างข้อมูลเสียงตัวเดิมกลับมา ดังในรูป (2.4)



รูป 2.4 (a) ระบบกำเนิดเสียงจริง (b) ระบบที่เลียนแบบการกำเนิดเสียงจริงและการวัดค่าความผิดพลาด

2.4 การเข้ารหัสเสียงแบบต่างๆ

การเข้ารุหัสเสียงแบ่งได้เป็น 2 พวกใหญ่ ๆ คือการเข้ารหัสเสียงตามรูปคลื่น(Waveform Coding)และการเข้ารหัสเสียงตามลักษณะของเสียงพูดหรือโวโคดเดอร์ การเข้ารหัสเสียงตามรูปคลื่นที่สำคัญที่จะยกตัวอย่างในที่นี้ได้แก่ PCM , DPCM และ ADPCM ส่วนโวโคดเดอร์ที่จะพูดถึงได้แก่ LPC-10 , CELP , LD-CELP และ VSELP ซึ่งจะพูดถึงหลักการทำงานโดยรวมดังนี้

Pulse Code Modulation (PCM) คือการเข้ารหัสเสียงแบบแรก ๆ เป็นตัวอย่างที่สำคัญของการเข้ารหัสเสียงแบบนอนพาราเมตริก เริ่มจากการสุ่มข้อมูลเสียงด้วยอัตรา 8 กิโลเฮิรตซ์ แล้วนำข้อมูลแต่ละตัวไปผ่านการควอนไตซ์เป็นข้อมูลดิจิตอลขนาด 8 บิตโดยใช้ตารางของการควอนไตซ์แบบลอการิธึม ทำให้ได้เป็น non-uniform PCM ที่มีอยู่ 2 แบบคือ A-law หรือ μ -law PCM ที่อัตราข้อมูล 64 กิโลบิตต่อวินาที มีอัตราส่วนของสัญญาณต่อเสียงรบกวน (signal-to-noise ratio) ประมาณ 83 dB เป็นการเข้ารหัสเสียงที่ใช้อ้างอิงเพื่อเปรียบเทียบกับการเข้ารหัสเสียงแบบอื่น ๆ และได้เป็นมาตรฐานของการเข้ารหัสเสียง G.711 ที่กำหนดโดย CCITT ในปีทศวรรษที่ 1960

ต่อมามีการดัดแปลงข้อมูลที่จะถูกควอนไตซ์ โดยแทนที่จะใช้ตัวข้อมูลโดยตรงก็จะใช้ผลต่างของข้อมูลตัวปัจจุบันกับข้อมูลที่ได้จากการทำนายพันธะเชิงเส้น นั่นคือจะทำการควอนไตซ์ $e(n) = s(n) - \hat{s}(n)$ โดยที่ $s(n)$ คือข้อมูลเสียงตัวปัจจุบันและ $\hat{s}(n) = \sum_{i=1}^{M} \hat{a}(i)s(n-i)$ คือเสียงที่ได้จากการทำนายพันธะเชิงเส้นโดยที่ $\hat{a}(i)$ เป็นค่าที่มีเก็บไว้ทั้งในเครื่องเข้ารหัสและเครื่องถอดรหัสเป็นวิธีการที่ใช้ใน **Differential PCM (DPCM)** และเมื่อมีการพัฒนาการปรับเปลี่ยนขนาดได้ทั้งค่าของขนาดขั้นในการควอนไตซ์และค่าข้อง $\hat{a}(i)$ จะได้เป็น **Adaptive DPCM (ADPCM)** ซึ่งมีการใช้อัตราเพียง 4 บิตต่อหนึ่งข้อมูล 32 กิโลบิตต่อวินาที ADPCM มีคุณภาพของเสียงที่ดีมากและเป็นมาตรฐานของการเข้ารหัสเสียง G.721 ที่กำหนดโดย CCITT ในปี 1988

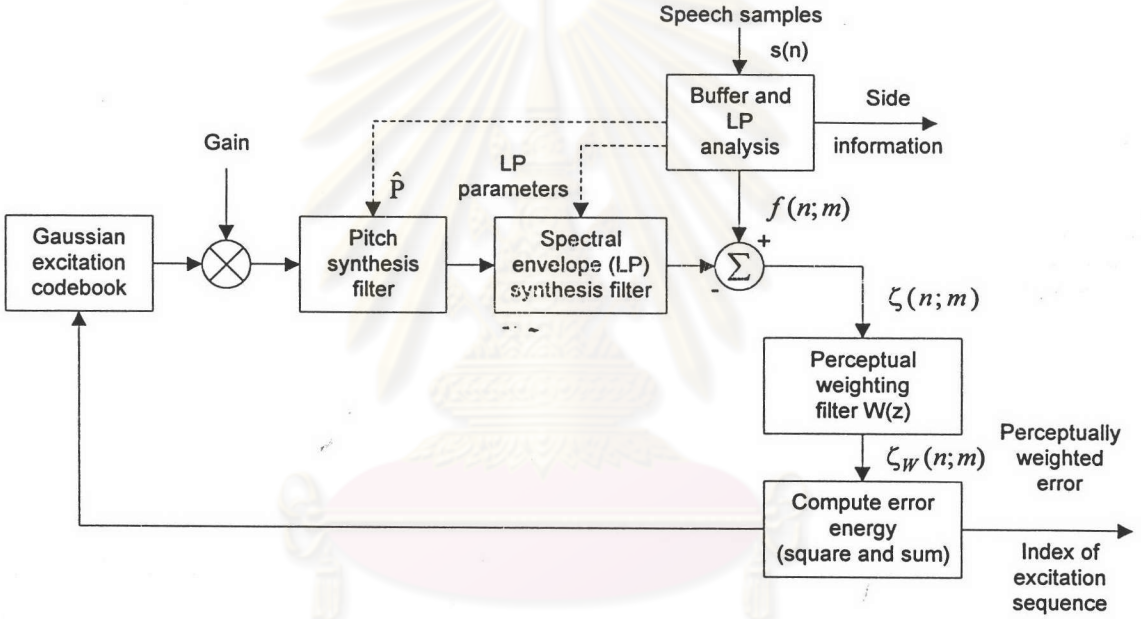
ต่อมามีรูปแบบการเข้ารหัสเสียงแบบต่าง ๆเกิดขึ้นมากมาย ซึ่งคงไม่สามารถกล่าวถึงในที่นี้ได้หมด ที่จะพูดต่อไปนี้จะเป็นเรื่องโวโคดเดอร์ซึ่งใช้หลักการกำเนิดเสียงที่เลียนแบบการกำเนิดเสียงพูดของมนุษย์ โดยเน้นที่เรื่องของโวโคดเดอร์แบบการทำนายพันธะเชิงเส้น(Linear predictive vocoder หรือ LPC)เนื่องจากเป็นเทคนิกที่มีการพัฒนาอย่างกว้างขวางที่สุดในช่วง 2 ทศวรรษหลังนี้

LPC-10 เป็นการเข้ารหัสรุ่นแรกที่ใช้เทคนิคของ LPC โดยการส่งลำดับของตัวกระตุ้นจะใช้ตัวกระตุ้นแบบขบวนของพัลส์ที่มีช่วงห่างกันเท่ากับคาบของเสียง(Pitch-pulse excitation)สำหรับเสียงก้องและใช้สัญญาณรบกวนสำหรับเสียงไม่ก้อง ซึ่งเป็นการกำหนดค่าลำดับของตัวกระตุ้นที่หยาบทำให้คุณภาพของเสียงที่ได้อยู่ในระดับเสียงสังเคราะห์ มีการคำนวณค่าสัมประสิทธ์ของการทำนายจำนวน 10 ตัวจึงมีชื่อว่า LPC-10

คุณภาพของเสียงที่ได้อยู่ในระดับเสียงสังเคราะห์(synthetic quality)เริ่มใช้ในของเล่นเพื่อการศึกษา " Speak and Spell " ของบริษัทเท็กซัสอินสตรูเมนท์ ในปี 1970s และเป็นมาตรฐานของการเข้ารหัสเสียง Federal Standard FS1015 ในปี 1984 การส่งข้อมูลระหว่างตัวเข้ารหัสและตัวถอดรหัสของ LPC-10 จะส่งค่าสัมประสิทธ์ของการทำนายและลำดับของตัวกระตุ้นไป ถ้าเป็นเสียงก้องก็จะส่งเฉพาะความสูงของพัลส์และคาบของเสียง ถ้าเป็นเสียงไม่ก้องก็จะส่งเฉพาะกำลังไปให้ทางด้านรับกำเนิดสัญญาณรบกวนเอาเองทำให้จำนวนข้อมูลที่ต้องใช้ต่ำ ทำงานได้ที่อัตราข้อมูลเท่ากับ 2400 บิตต่อวินาที

Code-Excited Linear Prediction (CELP) คือการเข้ารหัสเสียงแบบ LPC ที่ใช้การเข้ารหัสเวกเตอร์ ***e(n)*** ด้วยรหัสที่มีเก็บอยู่ในชุดรหัสจำนวนจำกัด โดยการทำงานจะทำกับสัญญาณเสียงที่เป็นเวกเตอร์ ให้คุณภาพของเสียงที่ดีโดยใช้อัตราข้อมูลที่ต่ำ CELP เป็นเครื่องเข้ารหัสที่มีการวิเคราะห์โดยการสังเคราะห์นั่นคือที่ด้านเข้ารหัสจะมีการสังเคราะห์เสียงตัวปัจจุบันขึ้นมาจากข้อมูลที่มีอยู่ แล้วนำไปเปรียบเทียบกับเสียงจริงที่เข้ามา ข้อแตกต่างที่ได้จะเป็นสัญญาณความผิดพลาด (Error signal)ที่ผ่านการถ่วงคุณภาพของการรับฟังแล้ว ซึ่งจะถูกนำไปผ่านการควอนไตซ์แบบเวกเตอร์เพื่อหาดัชนีของลำดับของตัวกระตุ้นที่เหมาะสมที่สุดแล้วส่งดัชนีนั้นไปยังด้านรับ นอกจากดัชนีแล้วยังต้องมีการส่งข้อมูลข้างเคียง(Side information)ไปตามช่องสัญญาณด้วย ข้อมูลเหล่านี้ได้แก่สัมประสิทธ์ของตัวทำนายและค่าคาบของเสียง

"CELP อาจจะเรียกได้ว่าเป็นตัวเข้ารหัสแบบไฮบริด(Hybrid Coder)เพราะว่ามีคุณสมบัติของทั้งการเข้ารหัสเสียงตามรูปคลื่นและการเข้ารหัสเสียงตามลักษณะของเสียงพูดรวมไว้ด้วยกัน" [7]



รูป 2.5 ตัวเข้ารหัสของ CELP แบบปกติ

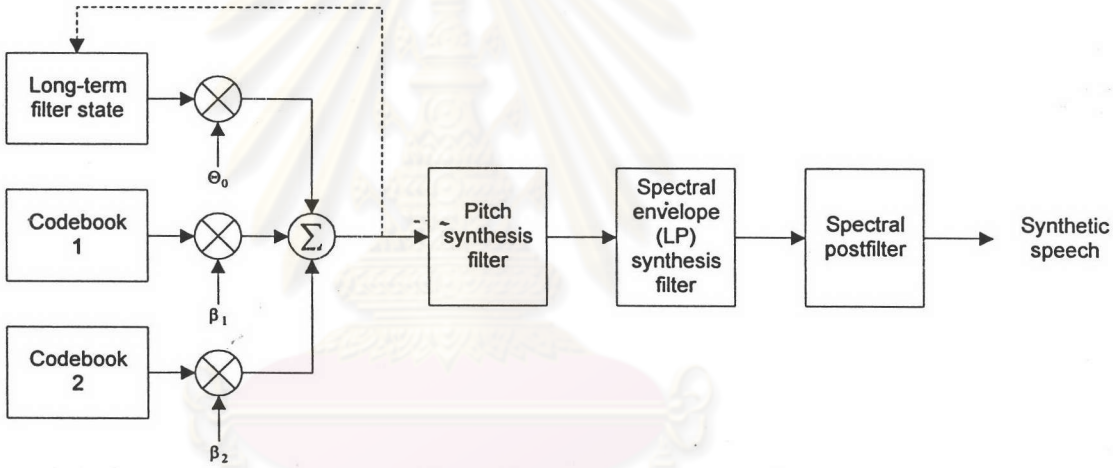
ในการควอนไตซ์แบบเวกเตอร์ใช้ชุดรหัสขนาด 10 บิตมีลำดับของตัวกระตุ้นได้ทั้งหมด 1024 แบบต่อเวกเตอร์ของตัวกระตุ้น(Excitation vector)ขนาด 40 ตัวอย่าง ซึ่งพบว่าเพียงพอสำหรับให้คุณภาพเสียงที่ดี CELP ที่ทำงานที่อัตราข้อมูล 16,000 บิตต่อวินาทีจะให้คุณภาพเสียงระดับเครือข่าย ซึ่งดีกว่าระบบอื่นที่ทำงานในอัตราเดียวกันแต่จะมีความล่าช้าทางเวลาที่มากกว่าเนื่องจากเป็นการเข้ารหัสที่ต้องใช้การวิเคราะห์ข้อมูลจำนวนมาก การทำงานเป็นลักษณะเฟรมต่อเฟรม ซึ่งขนาดของเฟรมที่ใช้จะเป็นเวลาประมาณ 40 ถึง 60 มิลลิวินาที แต่ยังมีวิธีที่จะลดความล่าช้าทางเวลาลงไปได้อีก ดังในข้อต่อไป

Low-Delay CELP (LD-CELP) ใช้อัตราข้อมูล 16 กิโลบิตต่อวินาที โดยให้คุณภาพของเสียงที่ดีเท่ากับหรือดีกว่าของ ADPCM แต่มีความต้องการการคำนวณที่สูงมาก ได้เป็นมาตรฐานของการเข้ารหัสเสียง G.728 ที่กำหนดโดย CCITT ในปี 1992 การทำงานโดยรวมเป็นการพัฒนามาจาก CELP แบบปกติแต่สามารถลดความล่าช้าทางเดียว (one-

way delay) ลงมาเหลือน้อยกว่า 2 มิลลิวินาที โดยการใช้การปรับเปลี่ยนค่าแบบย้อนหลัง(backward adaptation) ของสัมประสิทธ์ตัวทำนายพันธะเชิงเส้น,อัตราขยายและสัมประสิทธ์การถ่วงคุณภาพการรับฟัง และยังใช้ขนาดของเวกเตอร์ของตัวกระตุ้นขนาดเล็กเพียง 5 ตัวอย่างเท่านั้น รายละเอียดของ LD-CELP จะกล่าวถึงในหัวข้อ 2.5

Vector Sum Excited Linear Prediction (VSELP) ให้คุณภาพเสียงระดับการสื่อสาร โดยใช้อัตราข้อมูล 8 กิโลบิตต่อวินาที เป็นมาตรฐานของการเข้ารหัสเสียงของระบบโทรศัพท์เคลื่อนที่ดิจิตอลแบบเซลลูลาร์ที่ใช้ในทวีปอเมริกาเหนือ มีลักษณะของชุดรหัสที่แตกต่างไปจาก CELP คือ มีชุดรหัส 3 ชุด ชุดรหัส 1 และชุดรหัส 2 แต่ละชุดรหัสจะประกอบด้วย 128 คำรหัสที่เกิดจากการรวมกันแบบเชิงเส้นของ basis 7 ตัว แทนที่จะกำหนดมาเป็น 128 คำรหัสที่เป็นอิสระต่อกัน และ basis เหล่านี้จะมีการปรับค่าเพื่อลดค่า total perceptually weighted error ให้น้อยที่สุด.

long-term filter state เป็นชุดรหัสที่ประกอบด้วย 128 คำรหัสเช่นกัน โดยเป็นชุดรหัสแบบปรับเปลี่ยนได้ ผลลัพท์จากแหล่งกำเนิดตัวกระตุ้น(Excitation source)ทั้ง 3 นี้จะถูกคูณด้วยอัตราขยายซึ่งแยกอิสระกันของแต่ละตัว แล้วนำมารวมกันเพื่อผ่านตัวกรอง



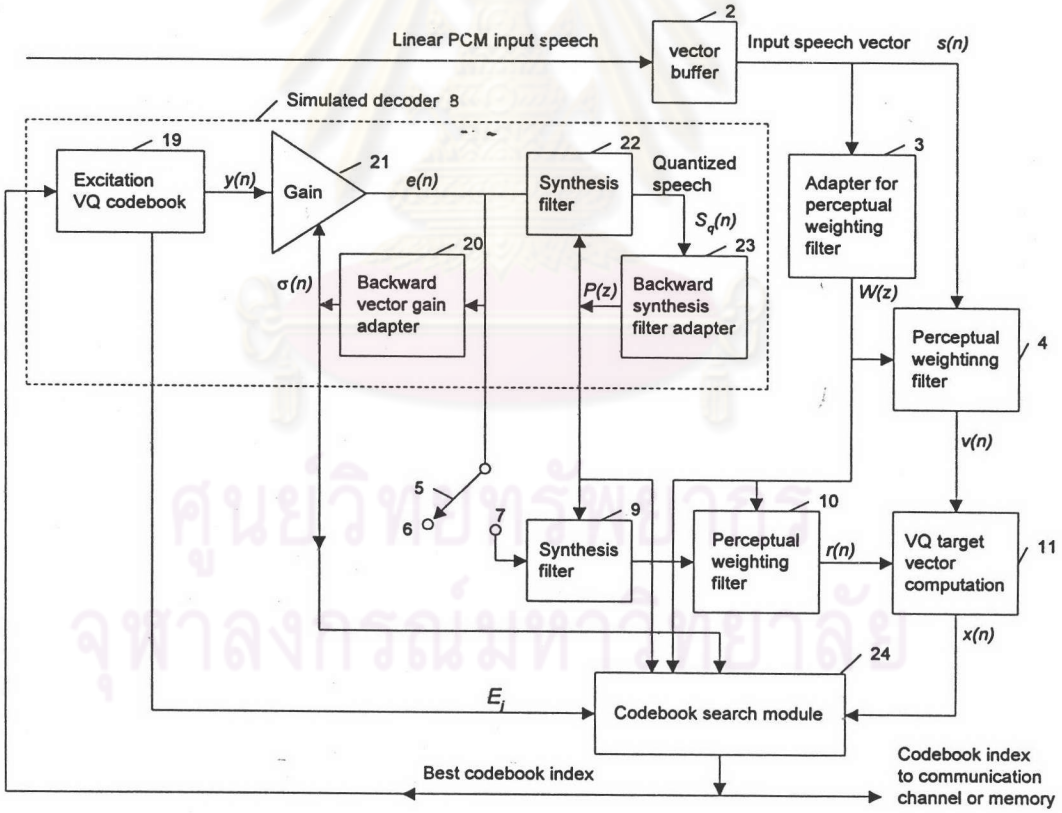
รูป 2.6 ตัวถอดรหัสของ VSELP

ในส่วนของตัวเข้ารหัสก็จะคล้ายกันกับตัวเข้ารหัสของ CELP โดยทั่วไป

2.5 มาตรฐาน ITU-T G.728 LD-CELP [2]

มาตรฐานของการเข้ารหัสเสียงพูดที่อัตรา 16 กิโลบิตต่อวินาทีโดยวิธีการทำนายพันธะเชิงเส้นแบบใช้รหัสโดยมีช่วงเวลาประวิงต่ำ(Low Delay Code Excited Linear Prediction หรือ LD-CELP)ถูกกำหนดขึ้นโดย CCITT ในปี ค.ศ.1992 ใช้เทคนิกการค้นหารหัสในชุดรหัสด้วยวิธีการวิเคราะห์จากการสังเคราะห์ซึ่งเป็นเทคนิกที่ใช้ในการเข้ารหัสเสียงพูดแบบ CELP โดยทั่วไป แต่ LD-CELP ใช้การปรับเปลี่ยนค่าสัมประสิทธิ์ของของตัวทำนายและอัตราขยายแบบย้อนหลัง ทำให้ได้การเข้ารหัสเสียงที่มีการหน่วงทางเวลาเพียง 0.625 มิลลิวินาที ในการส่งข้อมูลจะส่งเฉพาะดัชนีของรหัสที่ได้ไปตามช่องสัญญาณ สัมประสิทธิ์ของตัวทำนายจะถูกปรับเปลี่ยนเป็นรายคาบโดยใช้การวิเคราะห์แบบ LPC บนสัญญาณเสียงที่ถูกควอนไตซ์ออกมาก่อนหน้าสัญญาณเสียงตัวปัจจุบัน อัตราขยายจะถูกปรับเปลี่ยนจากข้อมูลของอัตราขยายที่อาศัยอยู่ในตัวกระตุ้นที่ถูกควอนไตซ์ออกมาเช่นกัน ขนาดของเวกเตอร์เสียงที่ใช้ 5 ตัวอย่างต่อ 1 เวกเตอร์ สัมประสิทธิ์ของวงจรกรองเพิ่มน้ำหนักการรับฟังของเสียง(Perceptual Weighting filter)จะถูกปรับเปลี่ยนโดยการวิเคราะห์แบบ LPC บนสัญญาณเสียงที่เข้ามา

2.5.1 ตัวเข้ารหัส(Encoder)



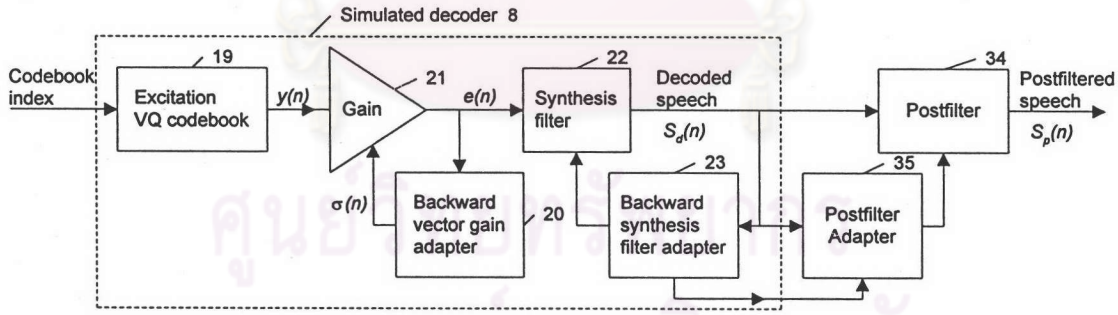
Encoder Block Schematic

รูป 2.7 บล็อกไดอะแกรมการทำงานของตัวเข้ารหัสของ LD-CELP

สัญญาณเสียงที่เข้ามาจะถูกแบ่งออกเป็นส่วน ๆที่เรียกว่าเวกเตอร์ โดยแต่ละเวกเตอร์นี้ประกอบด้วย 5 ตัวอย่างสุ่มที่อยู่ติดกัน ตัวเข้ารหัสจะส่งเวกเตอร์รหัสจากชุดรหัสขนาด 1024 ตัว ผ่านหน่วยปรับอัตราขยาย(Gain scaling unit) และวงจรกรองสังเคราะห์(synthesis fillter) จากชุดรหัส 1024 ตัวนี้ตัวเข้ารหัสจะระบุชุดรหัสที่ให้ค่าความผิดพลาดกำลังสองเฉลี่ยตามน้ำหนักของความถี่(Frequency-weighted mean-square error)ที่น้อยที่สุดเมื่อทำการวัดเปรียบเทียบกับเวกเตอร์ของเสียงที่เข้ามา ดัชนีชุดรหัสขนาด 10 บิตของเวกเตอร์รหัส(codevector)ที่ได้จะถูกส่งไปตามช่องสัญญาณไปยังตัวถอดรหัส(decoder) หลังจากนั้นเวกเตอร์รหัสนั้นจะถูกส่งผ่านหน่วยปรับอัตราขยายและวงจรกรองสังเคราะห์เพื่อปรับความจำของวงจรกรองใหม่และเตรียมการเข้ารหัสสัญญาณเสียงเวกเตอร์ต่อไป สัมประสิทธิ์ของวงจรกรองสังเคราะห์และอัตราขยายจะถูกปรับเปลี่ยนเป็นรายคาบ โดยวิธีปรับจากค่าย้อนหลังบนสัญญาณเสียงที่ถูกควอนไตซ์และตัวกระตุ้นปรับอัตราขยายก่อนหน้านี้ การใช้ดัชนีขนาด 10 บิตแสดงแทนเวกเตอร์เสียงขนาด 5 ตัวอย่างสุ่มหมายความว่าใช้อัตราข้อมูล 2 บิตต่อ 1 ตัวอย่างสุ่ม เนื่องจากอัตราการสุ่มข้อมูลเท่ากับ 8 กิโลเฮิรตซ์ จึงใช้อัตราการรับส่งข้อมูลเท่ากับ 16 กิโลบิตต่อวินาที

2.5.2ตัวถอดรหัส (Decoder)

เมื่อตัวถอดรหัสได้รับดัชนีขนาด 10 บิต ตัวถอดรหัสจะทำการเปิดตารางค้นหาเวกเตอร์รหัสที่ตรงกับดัชนีจากชุดรหัส และส่งเวกเตอร์รหัสที่ได้ผ่านหน่วยปรับอัตราขยายและวงจรกรองสังเคราะห์เพื่อสร้างเวกเตอร์สัญญาณเสียงที่ได้จากการถอดรหัสตัวปัจจุบันออกมา การปรับค่าของสัมประสิทธิของวงจรกรองสังเคราะห์และอัตราขยายใช้วิธีเดียวกับตัวเข้ารหัสเสียงที่ได้ต้องผ่านโพสต์ฟิลเตอร์เพื่อปรับปรุงคุณภาพในการรับฟัง สัมประสิทธิของโพสต์ฟิลเตอร์ถูกปรับค่าเป็นรายคาบโดยใช้ข้อมูลที่มีอยู่ในตัวถอดรหัสเอง เสียงที่ออกจากโพสต์ฟิลเตอร์ก็คือเสียงที่ได้จากการถอดรหัสโดยสมบูรณ์ รูปแสดงการทำงานของตัวถอดรหัสจะเหมือนกับในส่วนจำลองการถอดรหัสที่มีอยู่ในตัวเข้ารหัสและเพิ่มตัวโพสต์ฟิลเตอร์เข้าไป



Decoder Block schematic

รูป 2.8 บล็อกไดอะแกรมการทำงานของตัวถอดรหัสของ LD-CELP

2.5.3 หลักการสำคัญของตัวเข้ารหัส (Encoder principle)

1) ดัชนี k หมายถึงดัชนีของตัวอย่างสุ่มและการสุ่มตัวอย่างจะใช้ที่อัตรา 8 กิโลเฮิรตซ์หรือการสุ่มตัวอย่าง 1 ครั้งใช้เวลา 125 ไมโครวินาที

2) กลุ่มของตัวอย่างสุ่ม(sample)ที่อยู่ติดกัน 5 ตัวเรียกว่าเวกเตอร์ เช่นเวกเตอร์ของเสียง เวกเตอร์ของตัวกระตุ้น เป็นต้น ดัชนีของเวกเตอร์แทนด้วย n

3) กลุ่มของเวกเตอร์ที่อยู่ติดกัน 4 ตัว เรียกว่าเฟรม(adaptation cycles หรือ frame)

มีพารามิเตอร์ 3 ตัว ที่มีการปรับเปลี่ยนเป็นรายคาบได้แก่อัตราขยายของตัวกระตุ้น(excitation gain) สัมประสิทธ์ของวงจรกรองสังเคราะห์(synthesis filter coefficient) และสัมประสิทธ์ของวงจรกรองเพิ่มน้ำหนักการรับฟัง(perceptual weighting filter coefficient) ค่าของพารามิเตอร์เหล่านี้ได้จากการปรับเปลี่ยนแบบย้อนหลังโดยการคำนวณจากสัญญาณที่มีอยู่ก่อนสัญญาณตัวปัจจุบัน อัตราขยายของตัวกระตุ้นถูกปรับเปลี่ยนทุก ๆเวกเตอร์ ในขณะที่ สัมประสิทธ์ของวงจรกรองสังเคราะห์และวงจรกรองเพิ่มน้ำหนักการรับฟังจะถูกปรับเปลี่ยนทุก ๆเฟรม

ถึงแม้ว่าการปรับเปลี่ยนโดยรวมจะทำทุก ๆ 1 เฟรม แต่ขนาดของบัฟเฟอร์ยังคงเป็น 1 เวกเตอร์ทำให้มีการหน่วงทางเวลาที่ต่ำกว่า 2 มิลลิวินาที

2.5.4 หลักการทำงานโดยย่อ

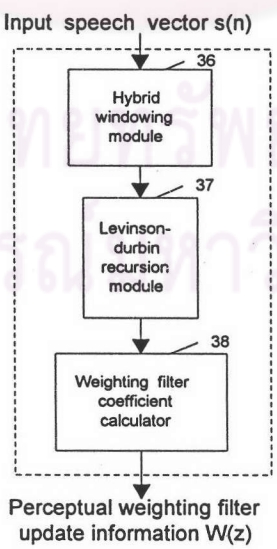
ข้อมูลเสียงขาเข้า(input speech) ตามมาตรฐานจะใช้เป็นยูนิฟอร์ม PCM ที่ถูกแปลงมาจาก μ-law หรือ *A*-law PCM แต่ในการทดลองนี้จะใช้ยูนิฟอร์ม PCM ที่ได้จาก A/D ที่มีมากับชุด DSK อยู่แล้ว จึงไม่ต้องเขียนบล็อกที่ 1 สำหรับการแปลงค่า PCM อีก รายละเอียดของ A/D อ่านได้ในบทที่ 3

เวกเตอร์บัฟเฟอร์(vector buffer) (บล็อกที่ 2)

จัดข้อมูลเสียงขาเข้า 5 ตัว ที่อยู่ติดกันเป็น 1 เวกเตอร์เสียง

ตัวปรับเปลี่ยนของวงจรกรองเพิ่มน้ำหนักการรับฟัง(adapter for perceptual weighting filter (บล็อกที่ 3)

ทำการคำนวณค่าสัมประสิทธิของวงจรกรองเพิ่มน้ำหนักการรับฟังค่าใหม่ทุก ๆ 4 เวกเตอร์ หรือ 1 เฟรมโดยใช้พื้นฐานการวิเคราะห์แบบการทำนายพันธะเชิงเส้นจากเสียงพูดที่ยังไม่ผ่านการควอนไตซ์ (เสียงพูดขาเข้า) ประกอบด้วย 3 ส่วนสำคัญดังนี้



รูป 2.9 บล็อกไดอะแกรมการทำงานของตัวปรับเปลี่ยนของวงจรกรองเพิ่มน้ำหนักการรับฟัง

บล็อกไฮบริดวินโดว์(Hybrid windowing module) (บล็อกที่ 36) คูณวินโดว์แบบไฮบริด(hybrid window)ซึ่งแสดงในรูป 2.10 เข้ากับเสียงพูดขาเข้าแล้วทำการคำนวณค่าอัตสหสัมพันธ์(Autocorrelation) 11 อันดับ เป็นค่า R(1) ถึง R(11)

บล็อกเลวินสัน-เดอร์บิน(Levinson-durbin recursion module) (บล็อกที่ 37) คำนวณสัมประสิทธิ์ของตัวทำนาย 10 อันดับแรกจากค่าของ R ที่ได้จากบล็อกที่ 36

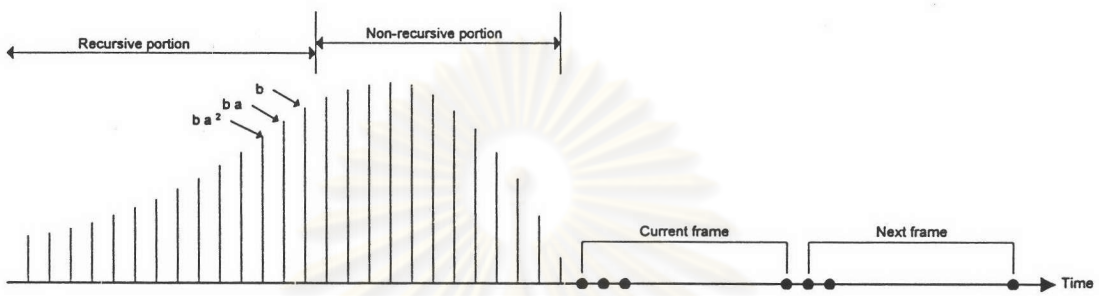


Illustration of a hybrid window

รูป 2.10 ลักษณะของไฮบริดวินโดว์

บล็อกคำนวณสัมประสิทธิ์ของวงจรกรองเพิ่มน้ำหนักการรับฟัง(Weighting filter coefficient calculator) (บล็อกที่ 38) คำนวณสัมประสิทธิ์ของวงจรกรองเพิ่มน้ำหนักการรับฟังค่าใหม่จากตัวทำนาย 10 ตัวที่ได้จากบล็อกที่ 37

วงจรกรองเพิ่มน้ำหนักการรับฟัง (บล็อกที่ 4)

เมื่อผ่านสัญญาณเวกเตอร์สัญญาณเสียงขาเข้าตัวปัจจุบัน ***s(n)*** เข้าไปจะได้เวกเตอร์เสียงที่ถูกปรับแต่งน้ำหนักแล้ว ***v(n)*** ออกมา

วงจรกรองสังเคราะห์ (บล็อกที่ 9)

เป็นวงจรกรองที่มีแต่โพล 50 อันดับ มีฟังก์ชันถ่ายโอนดังนี้

$$F(z) = 1/[\ 1- P(z)\ ] \qquad (2.9)$$

โดยที่ P(z) คือฟังก์ชันถ่ายโอนของตัวทำนาย LPC ขนาด 50 อันดับ

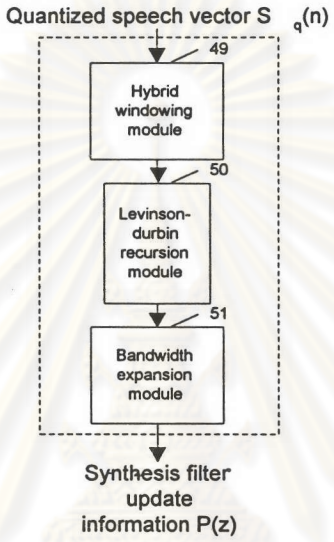
วงจรกรองสังเคราะห์(บล็อกที่ 9)และวงจรกรองเพิ่มน้ำหนักการรับฟัง(บล็อกที่ 10)จะคำนวณค่าเวกเตอร์ผลตอบเมื่อสัญญาณขาเข้าเป็นศูนย์(Zero-input response vector) ***r(n)*** ถึงแม้ว่าสัญญาณขาเข้าจะเป็นศูนย์ แต่ยังมีความจำ(memory)ของวงจรกรองอยู่ ทำให้ค่าของเวกเตอร์ที่ได้ไม่เป็นศูนย์ การปรับปรุงความจำของวงจรกรองจะเกิดขึ้นตลอดเวลาของการทำงาน

บล็อกคำนวณเป้าหมายของ VQ(VQ target vector computation) (บล็อกที่ 11)

คำนวณค่าเป้าหมายของการค้นหาชุดรหัส(VQ codebook search target)

$$x(n) = v(n) - r(n) \qquad (2.10)$$

ตัวปรับเปลี่ยนแบบย้อนหลังของวงจรกรองสังเคราะห์(Backward synthesis filter adapter) (บล็อกที่ 23)

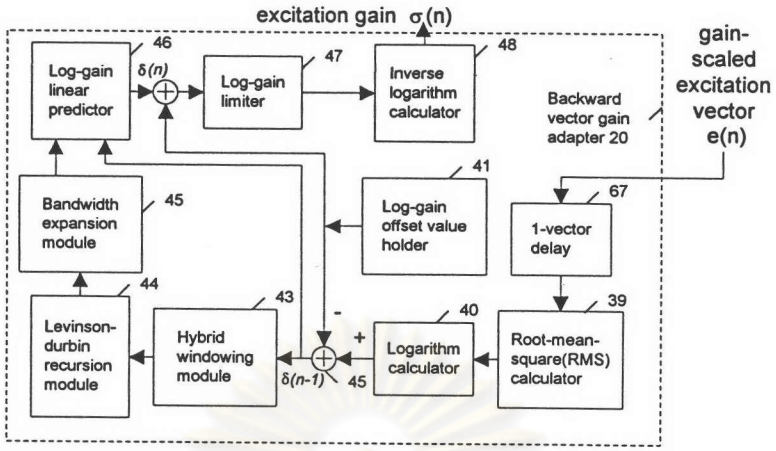


รูป 2.11 บล็อกไดอะแกรมการทำงานของตัวปรับเปลี่ยนของวงจรกรองสังเคราะห์

คล้ายกับบล็อกที่ 3 แต่ต่างกันตรงที่ใช้เสียงที่ผ่านการควอนไตซ์แล้วเป็นสัญญาณขาเข้า และคำนวณค่าอัตสหสัมพันธ์ถึง 51 อันดับ รวมทั้งสัมประสิทธิ์ตัวทำนาย 50 อันดับแทนที่จะเป็น 11 และ 10 อันดับตามลำดับอย่างในบล็อกที่ 3 บล็อกย่อยสุดท้ายของบล็อกนี้คือบล็อกขยายแถบความถี่(bandwidth expansion module)จะดึงโพลทุกตัวของวงจรกรองสังเคราะห์ที่คำนวณได้ให้เข้ามาใกล้จุดกำเนิดมากขึ้นด้วยอัตราส่วน λ ($\lambda < 1$) ในที่นี้ใช้ $\lambda = 253/256$ การเลื่อนโพลออกห่างจากวงกลมหนึ่งหน่วยมากขึ้นทำให้จุดยอดของผลตอบเชิงความถี่มีความกว้างมากขึ้น

ตัวปรับเปลี่ยนแบบย้อนหลังของอัตราขยาย(Backward vector gain adapter) (บล็อกที่ 20)

ตัวปรับเปลี่ยนตัวนี้จะปรับเปลี่ยนขนาดของอัตราขยายของตัวกระตุ้นทุก ๆ 1 เวกเตอร์ มีเวกเตอร์ตัวกระตุ้นที่ถูกขยายแล้ว(gain-scaled exication vector) e(n) เป็นสัญญาณขาเข้า และได้อัตราขยายของตัวกระตุ้น $\sigma(n)$ ออกมาโดยใช้การทำนายอัตราขยายของ e(n) จาก e(n-1) , e(n-2) , ... โดยการทำนายพันธะเชิงเส้นแบบปรับเปลี่ยนได้ (adaptive linear prediction) บนโดเมนของอัตราขยายแบบลอการิทึม ดังแสดงในรูป 2.12



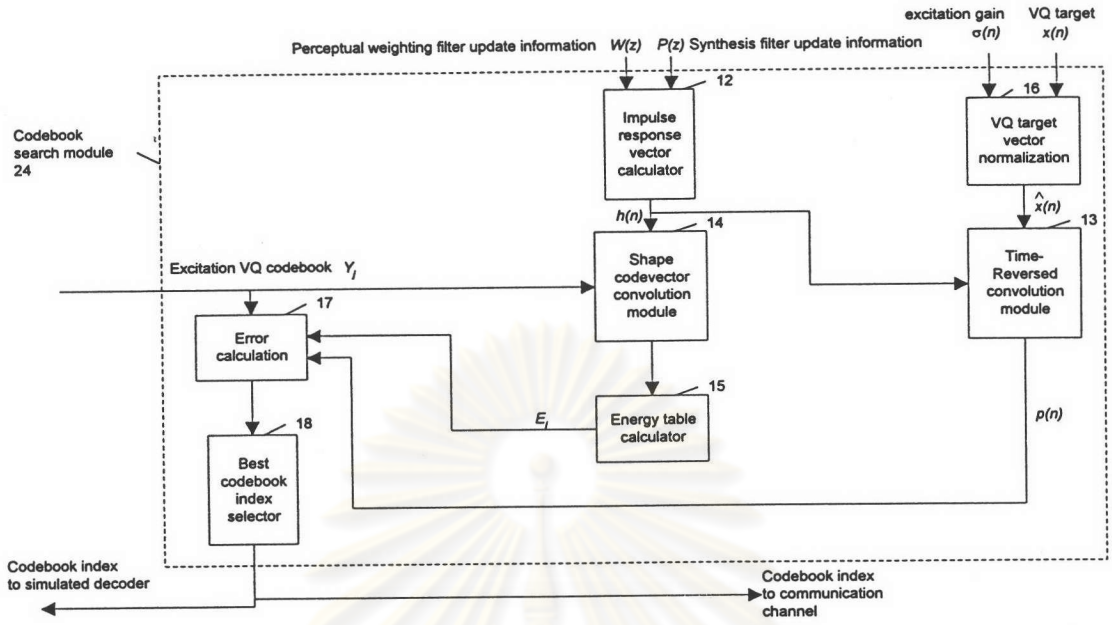
รูป 2.12 บล็อกไดอะแกรมการทำงานของส่วนปรับเปลี่ยนอัตราขยาย

ก่อนจะทำการทำนายค่า σ(n) จะต้องทำการเปลี่ยนค่า e(n-1) ที่ได้ให้เป็นโดเมนของลอการิทึมก่อนแล้วลบออกด้วยค่าออฟเซ็ต(offset)เพื่อให้ค่าที่ได้มีค่าใกล้เคียงกับศูนย์ โดยออฟเซ็ตที่ใช้เป็นค่าอัตราขยายแบบลอการิธึมของเสียงทั่วไป หลังจากนั้นจะไปผ่านบล็อกไฮบริดวินโดว์ บล็อกเลวินสัน-เดอร์บินและบล็อกขยายแถบความถี่ ซึ่ง 3 บล็อกนี้จะทำงานคล้ายกับบล็อก 23 ที่กล่าวก่อนหน้านี้จะมีแตกต่างกันที่ชนิดของสัญญาณขาเข้าและจำนวนอันดับของวงจรกรองที่ต้องคำนวณออกมา เมื่อได้ค่าสัมประสิทธิ์ของตัวทำนายแล้วก็จะทำนายอัตราขยายในรูปลอการิทึมออกมา จากค่าอัตราขยายแบบลอการิทึมที่ได้จะต้องบวกค่าออฟเซ็ตที่ลบออกไปก่อนหน้านี้กลับคืนเข้าไป ผ่านตัวจำกัดขนาด(limiter)เพื่อจำกัดขนาดไม่ให้ใหญ่เกินไปและขั้นตอนสุดท้ายคือการถอดค่าลอการิทึมออกก่อนเพื่อให้ได้อัตราขยายที่ต้องการนั่นคือ σ(n) โดย σ(n) จะถูกจำกัดไว้ที่ค่าระหว่าง 1 ถึง 1000

บล็อกการค้นหาชุดรหัส(Codebook search module) (บล็อก 24)

ประกอบด้วยบล็อก 12 ถึงบล็อก 18 ทั้งหมดร่วมกันคำนวณหาเวกเตอร์รหัสที่เหมาะสมที่จะให้เวกเตอร์ของเสียงที่ผ่านการควอนไตซ์แล้วที่ใกล้เคียงกับเวกเตอร์เสียงที่เข้ามามากที่สุด เพื่อลดประมาณการคำนวณลงในมาตรฐานจะแบ่งชุดรหัสออกเป็น 2 ส่วนคือชุดรหัสรูปร่างขนาด 7 บิต (จำนวน 128 เวกเตอร์รหัสที่เป็นอิสระต่อกัน) อีกส่วนคือชุดรหัสอัตราขยายขนาด 3 บิตจำนวน 8 ตัวที่สมมาตรเมื่อเทียบกับศูนย์ โดยจะแบ่งเป็น 1 บิตสำหรับเครื่องหมายและสำหรับขนาดอีก 2 บิต เมื่อเลือกเวกเตอร์รหัสได้ทั้ง 2 ส่วนแล้วนำมารวมกันก็จะได้ดัชนีของชุดรหัสขนาด 10 บิต

หลังจากได้ได้ดัชนีของชุดรหัสที่ต้องการแล้วก็จะทำการส่งดัชนีนั้นไปตามช่องสัญญาณเพื่อที่ตัวถอดรหัสจะได้สร้างสัญญาณเสียงกลับคืนมาได้ นอกจากนี้ดัชนีของชุดรหัสจะถูกส่งไปยังส่วนจำลองการถอดรหัส(simulated decoder) เพื่อสังเคราะห์เสียงพูดขึ้นมา เสียงที่สังเคราะห์ขึ้นมานี้จะเหมือนกับเสียงที่ทางด้านถอดรหัสจริง ๆสร้างได้ถ้าไม่เกิดความผิดพลาดขึ้นภายในช่องสัญญาณ เสียงสังเคราะห์นี้จะใช้เป็นองก์ประกอบในการเข้ารหัสเสียงพูดเวกเตอร์ถัดไปต่อไป ในการทดลองนี้จะไม่มีการส่งดัชนีของชุดรหัสไปตามช่องสัญญาณในส่วนของการทำงานตามเวลาจริง แต่จะเอาเสียงสังเคราะห์ที่สังเคราะห์ขึ้นมาที่ส่วนจำลองการถอดรหัสส่งออกมารับฟังกันโดยตรง



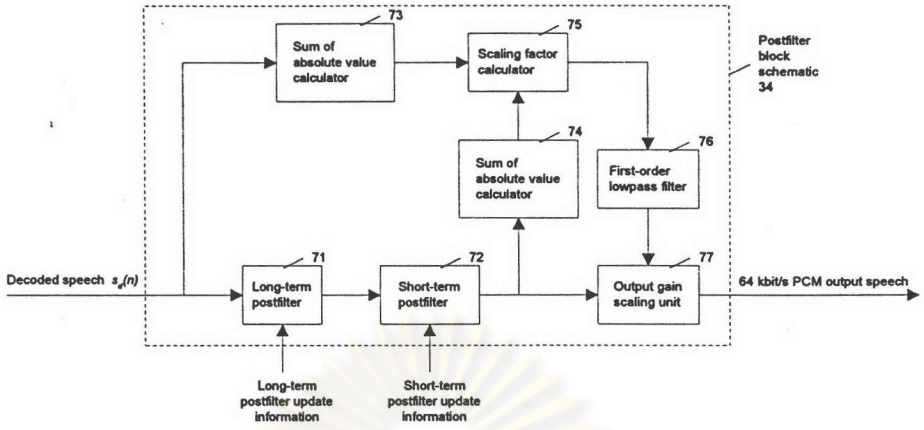
รูป 2.13 บล็อกไดอะแกรมการทำงานของการค้นหาชุดรหัส

โพสต์ฟิลเตอร์แบบปรับเปลี่ยนได้(Adaptive postfilter) (บล็อกที่ 34)

ใช้ปรับปรุงผลการรับฟังโดยการเน้นสเปกตรัม(spectrum)ของสัญญาณเสียงที่ถูกสังเคราะห์ขึ้นมาในช่วงจุดยอดของสเปกตรัม(spectral peak)และลดทอนลงในช่วงของห้วงระหว่างยอดสเปกตรัม(spectral valley)ในลักษณะที่คล้ายการทำงานของวงจรกรองแบบแมตช์(matched filter)โดยจะต้องคำนวณหาทั้งสเปกตรัมในช่วงยาว(long-term spectrum หรือ pitch)และสเปกตรัมในช่วงสั้น(short-term spectrum)เพื่อใช้ในการหาค่าสัมประสิทธิ์ของโพสต์ฟิลเตอร์และจะปรับปรุงค่าใหม่ทุก ๆ 4 เวกเตอร์หรือ 1 เฟรม

บล็อกโพสต์ฟิลเตอร์แบบปรับเปลี่ยนได้นี้จะมีอยู่เฉพาะด้านตัวถอดรหัสเท่านั้น เพื่อใช้เพิ่มคุณภาพของเสียงหลังจากถูกถอดรหัสให้สามารถผ่านการเข้ารหัสและถอดรหัสอย่างต่อเนื่องได้หลายครั้งโดยที่ยังรักษาคุณภาพของเสียงที่ดีเอาไว้ได้

ในการจำลองโปรแกรมบนตัวจำลองโปรแกรม(simulator)ด้วยภาษาแอสเซมบลีและการทำงานตามเวลาจริงเป็นการเขียนโปรแกรมเฉพาะส่วนของตัวเข้ารหัสไม่มีการเก็บหรือส่งดัชนีของชุดรหัสไปยังตัวถอดรหัสเพราะมีข้อจำกัดในด้านการรับส่งข้อมูล,เนื้อที่ในหน่วยความจำและเวลาในการทำงานจึงไม่มีการใช้บล็อกโพสต์ฟิลเตอร์แบบปรับเปลี่ยนได้ แต่การจำลองโปรแกรมบนแมทแลบไม่มีข้อจำกัดด้านหน่วยความจำหรือเวลาในการทำงานเพราะไม่ใช้การทำงานตามเวลาจริง จึงได้เขียนโปรแกรมของส่วนถอดรหัสรวมทั้งบล็อกโพสต์ฟิลเตอร์เอาไว้ด้วย



รูป 2.14 บล็อกไดอะแกรมการทำงานของโพสต์ฟิลเตอร์